

Similarity-Navigated Graph Neural Network: Energy-Efficient De-Duplication for Healthcare Data Aggregation in IoT Environments

Aishwarya Shekhar

School of Computer Science and Engineering, Galgotias University, Greater Noida, India
E-mail: aishwarya1.2.shekhar@outlook.com
ORCID iD: <https://orcid.org/0009-0001-7005-3760>

Abdul Aleem*

School of Computer Science and Engineering, Galgotias University, Greater Noida, India
E-mail: abdulaleem.cse765@gmail.com
ORCID iD: <https://orcid.org/0000-0001-6676-9072>

*Corresponding author

Received: December 27, 2024; Revised: May 20, 2025; Accepted: August 25, 2025; Published: August 8, 2026

Abstract: The rapid expansion of the Internet of Things (IoT) has enabled real-time patient monitoring through medical sensors, but redundant readings significantly increase storage, transmission, and energy costs. To address this challenge, we propose the Energy-Schoof's Cryptography-based Similarity-Navigated Graph Neural Network with Human Memory Optimization (ESC-SNGNNet-HMO) for healthcare data aggregation. The framework efficiently manages data chunks by combining deduplication, energy-aware processing, and secure transmission. At the fog layer, a Similarity-Navigated Graph Neural Network (SNGNN) identifies duplicate records, with hyperparameters optimized through Human Memory Optimization (HMO) to enhance accuracy. Deduplicated data is then securely transferred to the cloud using Schoof's Dynamic Elliptic Curve Cryptography (SDECC). Experimental evaluation demonstrates that ESC-SNGNNet-HMO maintains a throughput of 230 KB/s even with 8% packet loss, reduces storage to as little as 14 bytes, and eliminates up to 99% of duplicate data in low-node scenarios. Overall, the system provides an energy-efficient and cyber-secure solution for redundancy management in IoT-based healthcare applications.

Index Terms: De-duplication, Energy Transformer, Fog Server, Human Memory Optimization, Schoof's Dynamic Elliptic Curve Cryptography.

1. Introduction

The IoT is one of the evolutionary technologies that has brought about change and convenience in people's everyday lives, particularly in the fields of health and industry [1]. Many of which can sense a wide range of data, such as temperature, humidity, fire detection, and others, are embedded into the IoT [2]. These sensors aggregate data and transmit it to computer hubs, where fast decision-making is possible based on the analysis of the investigated conditions. This feature is most beneficial in the healthcare sector since wearable technology connected to IoT monitors patients' vital signs [3]. These gadgets include portable devices for remote patient condition assessment and can be worn over the skin surface, as accessories, or incorporated into the garments [4]. They ensure that medical personnel are always updated in their decision-making processes through the processing and transmission of health data to a central system [5]. In IoT-based systems, data aggregation is a functional undertaking that involves collecting and cumulating data into an intermediary node and then transmitting it. Based on the organization and management of the WSNs, the environment in IoT-based WSNs can be either homogeneous or heterogeneous [6]. What makes the homogeneous network reasonably easy to maintain is the fact that all of its nodes are the same: they all supply the same kind of information. Data aggregation procedures become more complex in a heterogeneous setting, on the other hand, when different nodes offer disparate types of data [7].

When dealing with a large network, several methods exist for properly handling data aggregation, as follows: Small networks often utilize CDA, where data is collected from all network nodes by an aggregator [8]. While CDA is easy to implement in small scenarios, it becomes a problem in transmission control and responsiveness in large

networks. On the other hand, several aggregator nodes are disseminated in the network using In-Network Data Aggregation (IDA) [9]. These nodes mitigate the drawbacks of CDA by collecting data from adjacent sensors and transmitting it to a Base Station (BS), thereby improving scalability and reducing transmission overhead. IoT-based systems require the de-duplication of data streams to compare and identify data replication, as the volume of data escalates significantly due to the proliferation of IoT devices. Two basic approaches in data de-duplication are inline processing and background processing [10,11]. Offline de-duplication occurs before data storage, where it compares the data against that already stored. In contrast, real-time de-duplication is processed during data storage and compares the data being stored against the existing data [12]. The most common problem created by redundancy in healthcare is the excessive storage of medical records, diagnostic reports, and patient histories, which consume space and energy, thereby affecting the network lifetime [13]. Energy consumption and the elimination of duplicative data collection are crucial strategies for preserving sensor node lifetimes and ensuring the reliability of health monitoring systems.

The techniques used in healthcare data aggregation and de-duplication face problems such as suboptimal aggregation, duplicate handling, high energy consumption, and high storage costs [14]. Effective solutions require the optimization of resource utilization, reduction of storage costs, and, simultaneously, the improvement of information transfer [15]. Healthcare IoT has potential that has not been seen before in improving patient care, enhancing system capabilities, and managing resources. To reduce redundancy, utilize resources effectively, and extend the network longevity in IoT environments, this research suggests an energy-efficient deep learning-based data aggregation system with an enhanced de-duplication technique.

Unlike existing approaches that primarily address either secure aggregation through cryptographic methods or data deduplication using machine learning, the proposed ESC-SNGNNet-HMO framework introduces a unified architecture that combines SNGNN with HMO for accurate and energy-efficient deduplication, and integrates this with SDECC to ensure secure data transmission. This joint design enables simultaneous improvements in redundancy elimination, storage reduction, energy efficiency, and cybersecurity within IoT healthcare systems. This work represents the first integration of GNN-based deduplication, optimization, and elliptic curve cryptography within a multispecialty IoT healthcare environment, establishing both methodological novelty and application-specific relevance. The key contributions are listed below.

- Development of a novel ESC-SNGNNet-HMO framework for energy-efficient healthcare data aggregation and deduplication in IoT environments.
- Design of an Aggregator Node that extracts relevant patient data using an Energy Transformer (ET) and forwards the processed information to the fog layer.
- Implementation of Similarity-Navigated Graph Neural Network (SNGNN) at the fog layer for deduplication, with hyperparameter optimization performed using Human Memory Optimization (HMO) to improve accuracy.
- Secure transfer of deduplicated data from the fog layer to the cloud server using Schoof's Dynamic Elliptic Curve Cryptography (SDECC), ensuring data integrity and confidentiality.

The research study is organized as follows: As introduced in Section 2, there has been a growing interest in developing models for this topic over the past few years. A further description of the recommended technique is provided in Section 3, while Section 4 presents the results of applying the improved strategy. Section 5 also presents the conclusion of the study.

2. Literature Survey

The list of recent papers on de-duplication for healthcare data aggregation is provided below,

In 2024, Saleh M. Aldowaijr [16] suggested an efficient data aggregation approach called Grid Hashing Efficient Data Aggregation (GH-EDA) for fast aggregation of IIoT systems, avoiding redundancy. In combination with preprocessing, data aggregation, and region splitting, the strategy provides a simple means of deduplication, reduction of data redundancy, division of the data by region, and fewer comparisons to identify duplicates using the Extended Merkle Grid. Standard sensor readings are transmitted in boolean form, which reduces processing costs, and a single hash calculation is performed on the data. While the system escalates work throughput and resource usage, it has limitations in dealing with highly heterogeneous data maps and dynamic overlay structures due to its reliance on zones.

In 2024, Mohd Akbar et al. [17] presented the cross-domain authentication to enforce secure data deduplication in cloud storage, the CB-TTDD approach. To avoid congestion in communication and storage, the Multi-Level Hashing Technique was adopted to enhance the deduplication technique and eliminate data copies simultaneously. Additionally, it can address other issues, such as boundary shifting, in addition to utilizing dynamic variable windows to enhance chunking efficiency. The system also performs parallel classification and aggregation's functionality for multi-class malware classification. Due to higher deduplication ratios, throughput, and processing speed, the proposed method is the safest and most efficient for distributed storage and big data compression.

In 2023, T. Benil and J. Jasper [18] introduced a novel blockchain-driven secure data outsourcing scheme for handling Electronic Health Records (EHR). The details of every activity are recorded in the technology's blockchain as

a transaction. Keyword searching ensures secure access, while registration guarantees the authenticity of each transaction. The method of encrypting EHRs while maintaining their confidentiality is by applying the Enhanced Map Chaotic Encryption (EMCE) technique. The process of data deduplication, which involves removing duplicate chunks, is performed using the Tri-Level Chunk Hashing (TL-CH) method. A TPA verifies the authenticity and provenance of the extracted EHRs. An appraisal of the scheme's operation demonstrates that it can effectively lower storage usage and mitigate standoff attacks.

M. Pavithra et al. [19] put forward the Boneh Goh Nissim Bilinear Attribute-based Optimal Cache Oblivious (BGNBA-OCO) solution to address secure deduplication and access control in cloud computing in 2024. This technique applies Boneh, Goh, and Nissim's Privacy-Preserving Revocable Attribute-Based Encryption to provide detailed access management with a reduced computational load by addressing attribute revocation and protecting sensitive data leakage. The optimal cache-oblivious algorithm helps to maintain a user's privacy by hiding the access patterns. The described method simplifies access policies and encrypted data changes, reducing the encryption overhead and duplication.

Khan et al. [20] proposed an Energy-Efficient Fuzzy Data Aggregation System (EE-FDAS) that could improve current sensor data processing in 2024. In normal readings and critical readings, the sensors are distinguished. Due to limitations on energy and available space for data storage, normal readings are summarized in a single Boolean digit, while vital values are sent in their formatted form. Data is summarized and sent to CHs for storage in the cloud, to be forwarded to Fog servers for physician access in the future. A comparative analysis demonstrated that the proposed EE-FDAS solution is effective in enhancing aggregation efficiency, reducing energy costs, and minimizing storage and transmission costs after all simulation procedures were completed.

In 2023, Zhiqiang Wang et al. [21] proposed a secure data-sharing scheme for cloud-aided EMSs, focusing on data reduction and concealing sensitive data. The technique masks Patient Identifying Information in EMRs with asterisks before encryption for privacy purposes and to optimize deduplication ratios. For the papers to be accessed by researchers, they are encrypted in a way that prevents anyone's identity from being revealed. Diagnostic data can be categorized according to duplicate ratios, and researchers can obtain the diagnostic data only when they need it. This method provides secure transfer and storage of electronic medical data, restricting its availability to databases and servers while striking a balance between data security, data deduplication, and data used for healthcare data management.

In 2022, Amir Javadpour et al. [22] proposed two dynamic models with different structures for mitigating DD and DR issues in BDC CNs. The models aim to improve performance by addressing certain issues present in earlier methods. While clustered architecture enhances reaction speed and load distribution, peer-to-peer enhances the use of the cloud network. The characteristics of a geographical area include bandwidth, replication frequency, and reputation change. Storage node attributes include rates, latency, failure probability, workload, and capacity. For that reason, simulations have better runtime performance results than simpler techniques; it is effective for a linear amount of time and requires less computational overhead.

Table 1. Summary of literature review.

Reference	Methods	Advantage	Limitations
[16]	GH-EDA	Lowers processing costs, minimizes data duplication, and effectively finds duplicates.	Reliance on predetermined zones limits scalability; it might not be able to manage shifting topologies or extremely varied data.
[17]	CB-TTDD	Decreases communication traffic, increases the effectiveness of deduplication, and strengthens security by using cross-domain authentication.	Depends on Windows for dynamic variables, which might make data management and chunking more difficult.
[18]	EMCE	Protects against intrusions, reduces storage needs, preserves confidentiality, and guarantees safe access.	The expense of encryption and blockchain integration may make it difficult for broad adoption.
[19]	BGNBA-OCO	Eliminates duplication, improves privacy, and offers fine-grained access control.	A computationally demanding encryption method that could not work well on devices with limited resources.
[20]	EE-FDAS	Lowers energy consumption, boosts aggregation efficiency, and lowers transmission and storage expenses.	It might not be capable of managing extremely varied or uncertain data situations.
[21]	MLE	Balances privacy and usability, guarantees safe exchange and restricted access and improves deduplication.	Reliant on the use of wildcards for sensitive data, which might cause problems with accuracy or precision.
[22]	Duplicate data model with peer-to-peer network	Increases speed, boosts performance and effectively balances load.	Restricted to certain cloud network circumstances; it might not be applicable in all circumstances.
[23]	SCDD-OVFKS	Optimizes key generation to increase security, boost efficiency, and use less storage.	The encryption and search techniques may result in processing overhead.

To integrate increased security and better storage efficiency in cloud data, Mohanaprakash Thottipalayam Andavan et al. [23] proposed a Secure Cloud Data Deduplication (SCDD) method in 2024. The technique ensures optimum security by implementing the Refraction Learning-based Chimp Optimization (RL-CO) algorithm for key generation

and a proxy re-encryption approach for the data. By using an Optimal Verified Fuzzy Keyword Search (OVFKS) method, duplication of files is eliminated, which significantly improves storage usage. The effectiveness and efficiency of the proposed SCDD-OVFKS technique were assessed when applied to different data types, including music and application files, compared to the other strategies. It demonstrated lower computational overhead, memory usage, and search time, but achieved a higher deduplication ratio. Table 1 provides a summary of the literature review.

Although several methods, such as GH-EDA, CB-TTDD, EMCE, BGNBA-OCO, EE-FDAS, SCDD-OVFKS, and other cloud or blockchain-based schemes, have achieved progress in reducing redundancy, improving security, or lowering energy consumption, most existing approaches face challenges in handling heterogeneous IoT healthcare data, ensuring scalability across multispecialty environments, and simultaneously achieving high deduplication accuracy with lightweight computation. Limited attention has been given to integrating graph-based learning with cryptographic and optimization techniques for real-time healthcare data deduplication and energy-efficient transmission.

2.1. Problem statement

The frequent transmission of repetitive sensor readings in IoT healthcare networks results in increased energy consumption, storage overhead, and communication costs. Existing approaches provide partial solutions, but they often suffer from scalability issues, computational complexity, or a lack of adaptability to heterogeneous healthcare data. Therefore, there is a need for a framework that can efficiently deduplicate patient data at the edge and fog layers while ensuring secure transmission, low energy usage, and reduced storage requirements in multispecialty IoT healthcare environments.

3. Proposed methodology

The proposed ESC-SNGNNet-HMO efficiently manages data chunks to reduce redundant transfers, minimize duplication, lower storage costs at central repositories, and decrease energy consumption in aggregator nodes. Patient data is continuously collected by medical sensor nodes and transmitted to an Aggregator Node, where an Energy Transformer (ET) extracts relevant information before forwarding it to the fog layer. At the fog layer, SNGNN performs deduplication, with hyperparameters optimized using HMO to improve detection accuracy. The deduplicated data is then securely transmitted to the cloud server through SDECC. The overall ESC-SNGNNet-HMO architecture is illustrated in Figure 1.

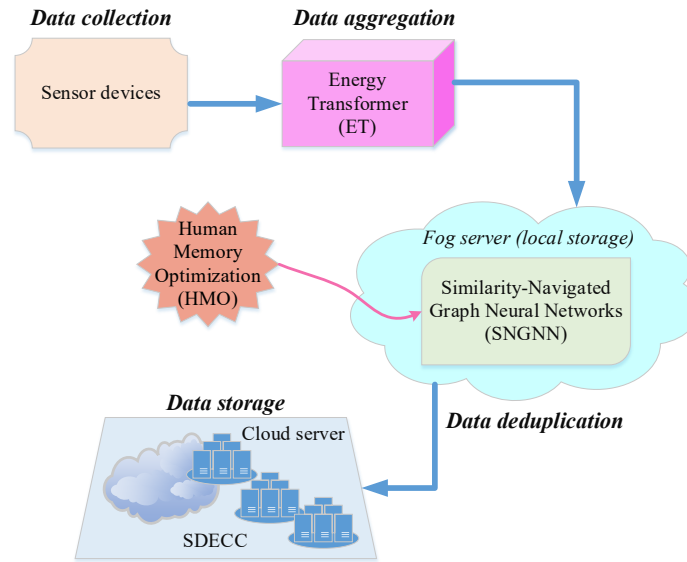


Fig. 1. Proposed ESC-SNGNNet-HMO architecture.

3.1. Data aggregation

The medical sensor nodes are always monitoring and acquiring patient information that they relay to the aggregator node for processing. The Energy Transformer (ET) [24] is used for data aggregation, which dynamically aligns and consolidates incomplete patient data before sending it to the Fog Server. Every patient's data record, including vital signs, results, and activity records, is considered a "token" in the ET framework. Each dimension corresponds to a sensor reading in the multi-dimensional space $y \in S^E$ where these tokens are embedded. The open tokens (y^B) represent complete and tagged sensor data. Tokens with masks (y^C) are used to represent incomplete data.

To stabilize the feature space and ensure consistent contributions from each sensor dimension, the input tokens are normalized. The normalization process is mathematically expressed in equation (1).

$$h_j = \gamma \frac{y_j - \bar{y}}{\sqrt{\frac{1}{E} \sum_k (y_k - \bar{y})^2 + \varepsilon}} + \delta_j \quad (1)$$

where, γ and δ_j are learnable parameters in this instance. The mean is denoted as $\bar{y} = 1/E \sum_{l=1}^E y_l$. The division by zero ends with a small constant denoted ε . Patient data streams exchange information with each other via energy-based attention. After normalization, the model aligns incomplete tokens with complete tokens using an energy-based attention mechanism. The attention score between tokens is defined in equation (2),

$$F^{atte} = -1/\beta \sum_{i=1}^I \sum_{D=1}^O \log(\sum_{C \neq D} \exp(\beta B_{iCD})) \quad (2)$$

The actual attention matrix B_{iCD} is computed in equation (3),

$$B_{iCD} = \sum_{\alpha} L_{\alpha iC} R_{\alpha iD} \quad (3)$$

The key matrix $L_{\alpha iC} = \sum_k X_{\alpha i k}^L h_k C$ and query matrix $R_{\alpha iD} = \sum_k X_{\alpha i k}^R h_k D$ are computed for tokens C and D ,

respectively, using the learnable parameters $X^L \in S^{Z \times I \times E}$ $X^R \in S^{Z \times I \times E}$. These matrices make it possible to match queries with the right keys, which are features that are available from other tokens. To identify missing data features, correlations with other tokens are utilized, ensuring dependable and contextually aware data aggregation. To maintain temporal consistency, the Hopfield Network (HN) module verifies token patterns against stored representations. The energy function of the Hopfield module is given in equation (4).

$$F^{HN} = - \sum_{C=1}^O \sum_{\mu=1}^L H \left(\sum_{k=1}^E \xi_{\mu k} h_k C \right), \quad \xi \in S^{L \times E} \quad (4)$$

$\xi_{\mu k}$ are learnable weight vectors. H Represents the integral of the activation function S . Equation (5) describes the joint minimization process, where the differential equation is discretized using the Euler method for efficient updates during training.

$$\tau \frac{dy_{jB}}{dt} = - \frac{\partial F}{\partial h_j B}, \quad F = F^{atte} + F^{HN} \quad (5)$$

This differential equation is discretized to enable efficient updates during aggregation. The attention term aligns data streams (tokens) by minimizing query-key mismatches. To regularize data consistency, the Hopfield term compares to previously learned patterns. The dynamics converge to a fixed-point attractor, resulting in imputed and aggregated patient data in the masked tokens (y^B). To verify the data stream is complete, the ET block sends it to the Fog Server for additional processing.

3.2. Data de-duplication using Similarity-navigated graph neural networks

Utilizing data similarities, Similarity-Navigated Graph Neural Networks (SNGNN) [25] efficiently identify and remove redundant data points in de-duplication. In the initial step, the Fog Server establishes a prevailing chunk size for the incoming data. This ensures consistency in data processing and prepares the data for efficient similarity comparison.

3.2.1. Node Embedding and Similarity Calculation

The data chunks are shown as nodes in a graph $I = (X, \varepsilon)$, which X stands for the data chunks and G potential connections between the chunks. For every node x_k , the n^{th} layer of the SNGNN creates embeddings $j_k^{(n)}$. These embeddings capture both the structural and feature information of the node. The data chunks are represented as nodes in a graph. To capture feature and structural similarity, each node embedding is generated by the GNN. The similarity between two embeddings is computed using cosine similarity, as given in equation (6).

$$R_{kl}^{(n)} = \frac{j_k j_l^V}{\|j_k\|_2 \|j_l\|_2} \quad (6)$$

where the Node Similarity Matrix is represented as $R_{kl}^{(n)}$, the cosine similarity between the embeddings of nodes x_k and x_l is determined by $R_{kl}^{(n)}$.

3.2.2. Duplicate data Identification

To determine a Linked Node Similarity Matrix $N^{(n)}$, which emphasizes important similarities, the adjacency matrix C , given by equation (7), is utilized.

$$N^{(n)} = R^{(n)} \otimes C \quad (7)$$

where multiplication by elements is represented by $\otimes$. The similarity between nodes that are directly connected is shown in this matrix. The Neighborhood Similarity Ranking is obtained by sorting the non-zero elements of the k^{th} row $N^{(n)}$ for each node x_k . This rating facilitates the process of prioritizing neighbors with the highest similarity scores. The deduplication process is enhanced with the use of explicit selection. Only the top m neighbors who are equivalent above a threshold θ are merged, as indicated in equation (8).

$$P_k^{(n)} = \{x_l: l \in \text{top } m \text{ neighbors of } x_k, R_{kl}^{(n)} > \theta\} \quad (8)$$

The node attributes x_k are aggregated using equation (9).

$$j_k^{(n)} = \frac{1}{|\bar{P}_k^{(n-1)}| + 1} \sum_{x_l \in \bar{P}_k^{(n-1)} \cup \{x_k\}} R_{kl}^{(n-1)} j_l^{(n-1)} \quad (9)$$

By limiting the influence of irrelevant data, this method guarantees that only important neighbors contribute to the feature propagation. To balance information between local node attributes and global graph structure, the fusion mechanism in equation (10) is applied.

$$J_{SNGNN++}^{(n)} = (1 - \alpha) \cdot J_{SNGNN+}^{(n)} + \alpha \cdot J_{mlp}^{(n)} \quad (10)$$

where, α is a trainable parameter. The deduplication capabilities are enhanced in this fusion by utilizing both the local node attributes and the global graph structure. The Fog Server efficiently identifies redundant chunks by utilizing the similarity scores and improved embeddings.

3.3. Hyperparameter optimizing using Human memory optimization algorithm

The HMO algorithm was inspired by human memory processes, which derive behavior from past experiences [26]. To find the optimal solution to an optimization problem, the approach simulates the processes of memory formation, recall, and forgetting. The HMO algorithm is explained below:

Step 1: Initialization: The population of individuals (solutions) is initialized randomly within the limits of the problem space. Each person represents one possible answer. dim is the dimensionality of the problem and P is the population size, the memory matrix's initial capacity $P \times \text{dim}$. The matrix contains recollections of both good and bad events. The worst solution, the greatest solution, and the average position $\tilde{Y}$ are examples of global variables that are initialized.

Step 2: Memory Production (Human Activity): In the memory production step, each individual solution updates its position in the search space according to equations (11)–(13)

$$Z_k(v+1) = \tilde{Z} + \forall \times (wd - nd) \times \text{rand}(1, \text{dim}) \quad (11)$$

$$\forall = m \times \sin(1/m) \times d \quad (12)$$

$$d = 2 \times \left(\frac{1-v}{V} \right) \quad (13)$$

Where, $\tilde{Z}$ is the population's average position. The problem space's upper and lower boundaries are denoted by wd and nd . The flexibility in exploring the solution space is represented by the divergence factor. d is a factor that diminishes with time, while m is a random constant between (0.1, 1.3). The maximum iteration count is V .

Step 3: Storing Memories (Successful and Failed Events): The greatest (successful) and worst (failing) memories are saved for each cycle. These memories are stored in a $P \times \text{dim}$ matrix. The newest memory takes the place of the oldest if it is larger than the matrix's capacity. Only better and worse answers than the prior ones are stored in the matrix. To track the algorithm's best and worst experiences, the global optimum and worst solutions are maintained.

Step 4: Recall Behaviour (Failed Memories): The recall process updates candidate solutions based on successful and failed memories. Equation (14) governs updates using failed memories,

$$Z_k(v+1) = z_{best} + \beta \times (t_1 \times worst_c - t_2 \times \tilde{Z}) \quad (14)$$

Where the best solution in the world is z_{best} . $worst_c$ is a bad memory chosen at random. With $\delta = 4$ and ϕ being a regularly distributed random vector, $\beta = -\frac{1}{\delta} \times \log(1-\phi)$ stands for the emotional state. The percentage of data taken from the poorest and average memories is represented by the random integers t_1 t_2 , which range from 0 to 1.

Step 5: Recall Behaviour (Successful Memories): while equation (15) governs updates using successful ones, which is expressed below,

$$Z_k(v+1) = z_{best} + \beta \times (t_1 \times best_{c2} - t_2 \times \tilde{Z}) \quad (15)$$

where, the memory $best_{c2}$ is chosen at random from the optimum memory matrix. The value β is unchanged from the memory recall failure. The algorithm focuses more on improving existing solutions in this stage, which highlights local exploitation, where outcomes become increasingly similar.

Step 6: Memory Loss (Transient Memory Loss): When a person has momentary memory loss, individuals utilize the following equation to update their position (16).

$$Z_k(v+1) = Z_t(v) + \sigma \times (t_1 \times \tilde{Z} - R_{worst}) \quad (16)$$

Where, a randomly chosen member of the population is denoted by $Z_t(v)$. A random factor σ denotes discontent with the condition of memory at that moment. The worst memory is the most recent one. The program diversifies the search by retrieving forgotten memories through global investigation.

Step 7: Normal Memory Update: If no memory is lost, update using the most suitable solution available. If the memory is not lost, the update is found in equation (17).

$$Z_k(v+1) = z_{best} + t_2 \times (Z_n - Z_k(v)) \quad (17)$$

where, Z_k represents the best option at the moment. The current mood state, which directs the exploitative behavior, is represented by the random integer t_2 .

Step 8: Termination: Until the best solution is discovered or the maximum number of iterations V is reached, the algorithm iterates through steps 2 through 7. The HMO flowchart is described in Figure 2.

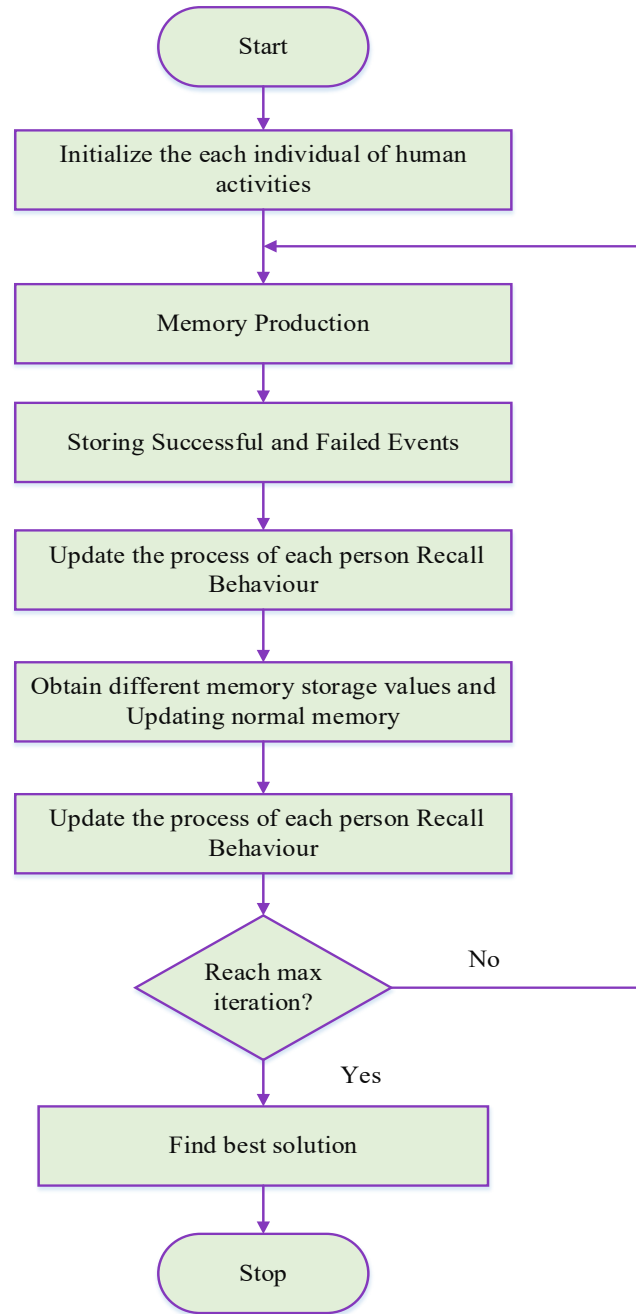


Fig. 2. Flowchart of HMO.

3.4. Data storage

To ensure secure transmission of deduplicated data, SDECC is applied [27]. The elliptic curve used is defined by equation (18),

$$z^2 = y^3 + by + c \quad (18)$$

where b , c , and the prime q is chosen at random using a secure Pseudo Random Number Generator (PRNG). To ensure that the curve is not unique, the discriminant $-16(4b^3 + 27c^2) \neq 0$ is verified. Elliptic Curve Discrete Logarithm Problem (ECDLP) resistance is increased by Schoof's Algorithm, which determines the number of points $\#F(G_q)$, on the curve in polynomial time while guaranteeing that it is a prime integer (O). Since the cofactor i is reduced to 1, the group is cyclic. Verifying that $iR = P_\infty$ and $oR = P_\infty$ allows for the selection of a safe generator point H . $public\ key = private\ key * H$ is the public key, and a private key ($private\ key \in [1, q-1]$) is selected. A message N is

mapped to a point on the curve via encryption, creating the ciphertext $D_N = \{Q_B H, (Q_N + Q_B * q_C)\}$, which q is a random scalar. After decryption, N is obtained as $Q_N = zD_N - Q_C y_{D_N}$. By dynamically producing safe key pairs and prime-order elliptic curves, this technique offers strong security and is appropriate for contexts with limited resources. The SDECC flowchart is described in Figure 3.

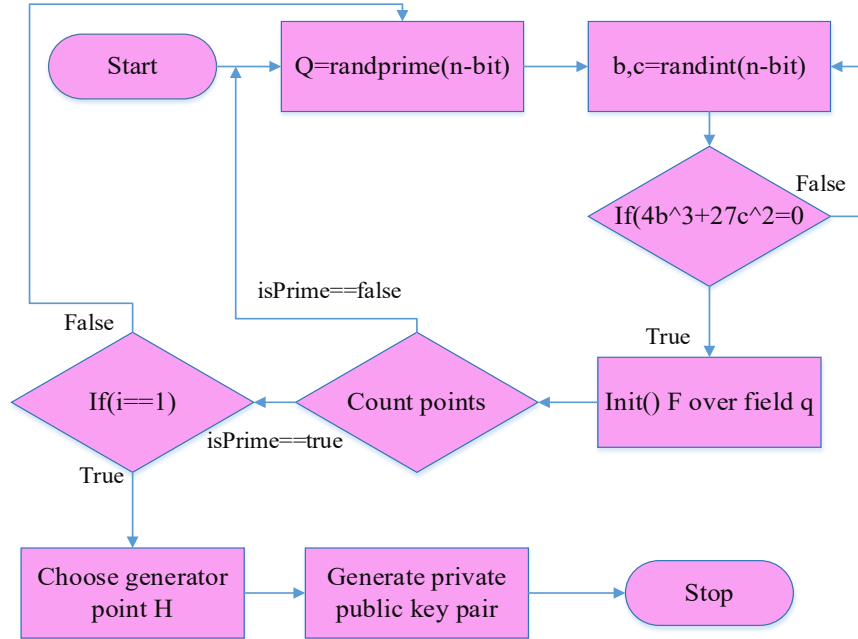


Fig. 3. SDECC flowchart.

4. Results and Discussion

The performance of the proposed ESC-SNGNNet-HMO system is evaluated in a 3000 x 3000 m network with a transmission range of 30 m using NS-2.35 simulations. With sensor nodes generating data packets every 30 seconds, the simulation runs for 2500 seconds. The tool command language is used when setting up node deployments, and the AWK scripts are used to obtain the processed results. The simulation parameters are presented in Table 2 below.

Table 2. Simulation parameters used for evaluating the ESC-SNGNNet-HMO framework.

Simulation parameters	Values
Network size	3000 x 3000 m
Simulation tool	NS 2.35
No of nodes	300
Nodes organization	Random
Receiving energy utilization (μJ)	0.049
Transmission range	30 m

4.1. Performance metrics

The performance assessment of the proposed ESC-SNGNNet-HMO system is presented in this section based on the energy consumption, average aggregation, throughput, packet loss ratio, deduplication ratio, and storage cost. To compare efficiency in both simple and diverse conditions, a heterogeneous environment with various characteristics is applied for simulations. These important factors enable the assessment of the scheme's ability to efficiently amalgamate and de-duplicate under various network performance conditions.

4.1.1. Deduplication ratio

The amount of storage space saved by implementing a deduplication solution is captured by a figure referred to as the deduplication ratio (DR). It demonstrates how effectively the system minimizes redundancy and allocates storage space efficiently. The deduplication ratio is given in equation (19).

$$DR = \frac{DS - DSS}{DSS} \quad (19)$$

here, DSS indicates the size of the data saved following deduplication, whereas DS indicates the overall size of the data before deduplication.

4.1.2. Throughput

Another important measure of the system's performance in handling data transmission is throughput, which measures how quickly deduplicated medical data is transferred to a cloud repository. It is described in equation (20).

$$Throughput = \frac{g_t}{avg_g} * \frac{1}{M_v} \quad (20)$$

where, M_v is the amount of time in seconds that network nodes need to transmit data, avg_g is the average size of each data transaction, and g_t is the overall size of the deduplicated data.

4.1.3. Average aggregation

The total amount of data packets gathered from sensor nodes before being sent to the Fog server is referred to as average aggregation. This measure, which evaluates the effectiveness of data collection in the system, is computed by running simulations five times and averaging the results to determine overall performance.

4.1.4. Energy consumption

Energy consumption (EC) is a critical factor in sensor networks since nodes have limited power. Energy is used for each bit transmitted during the data transmission phase. By compressing the data, the proposed technology maximizes energy consumption. The energy consumption is calculated using equation (21).

$$EC = \frac{Throughput}{Total\ power\ consumed} \quad (21)$$

4.1.5. Packet loss ratio

The packet loss ratio calculates the proportion of data packets lost during transmission. Larger packets are more likely to be lost due to network congestion and data corruption. The recommended technique reduces packet loss and ensures smaller packet sizes that can be transported efficiently, even with constrained bandwidth, by compressing data.

4.2. Performance analysis

The performance analysis proposed by ESC-SNGNNet-HMO is used to evaluate the effectiveness of the deduplication strategy in certain scenarios. To analyze important metrics such as performance, deduplication ratio, storage space, communication overhead, and encryption/decryption timings, various file sizes and numbers of files are presented. The analysis is apparent in terms of balancing encryption and decryption efficiency, reducing storage requirements, and boosting throughput.

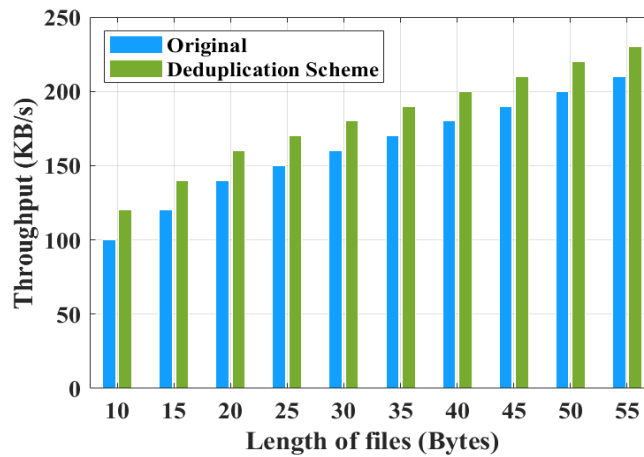


Fig. 4. Throughput comparison of the proposed deduplication strategy and baseline system across different file sizes.

Figure 4 compares the performance of a deduplication technique with that of the original system as the file size increases from 10 to 55 bytes. Initially, the deduplication technique outperforms the old scheme, achieving approximately 100 KB/s for 10-byte files, compared to 85 KB/s. Both techniques improve as the file size increases. Both strategies perform similarly at 40 bytes, achieving about 200 KB/s. Beyond this, the deduplication strategy

performs somewhat better than the original system, achieving 230 KB/s for files of 55 bytes in size, compared to the old scheme's 220 KB/s. This illustrates the benefit of the deduplication approach, especially for longer files.

Figure 5 shows how throughput (purple) and deduplication ratio (cyan) relate to the number of files (in thousands). Throughput begins at a high normalized value of 1.5 for 20k files and reduces progressively to about 1 for 100k files as the number of files grows from 20k to 100k. Likewise, the deduplication ratio consistently maintains a normalized value near 1 for all file counts. The findings show that the approach is effective in preserving storage optimization as the number of files increases, even while performance appears to decrease slightly with bigger datasets.

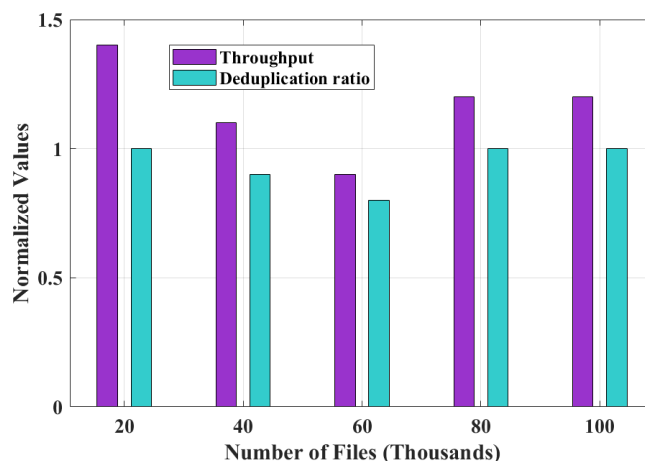


Fig. 5. Normalized throughput and deduplication ratio as a function of file count.

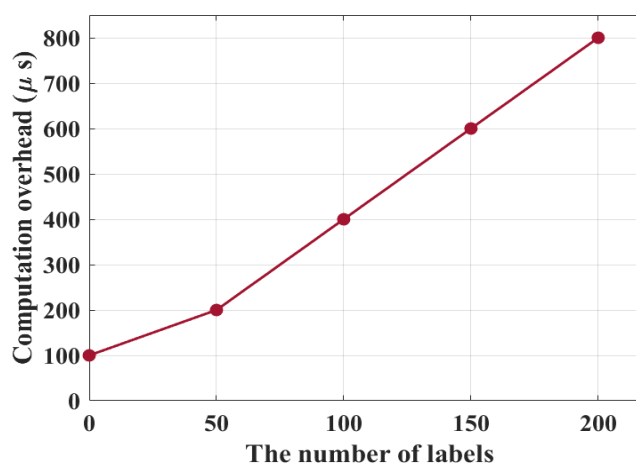


Fig. 6. Communication overhead with respect to the number of labels.

The link between the number of labels and communication overhead is seen in Figure 6. The transmission overhead rises linearly with the number of labels. When there are no labels, the overhead is almost 100 microseconds (μs), and it progressively grows as the number of labels rises, according to the plot. At fifty labels, overhead rises to around 200 μs , and at 200 labels, it surpasses 800 μs . Given that the communication overhead rises with the number of labels, this demonstrates how the number of labels may impact system performance.

The storage space requirements for non-deduplication, perfect deduplication, and deduplication are contrasted in Figure 7. Storage grows linearly with file size in the non-deduplication scenario, reaching about 500 GB for 100,000 files. Because all duplicates are removed, perfect deduplication maintains a consistent storage utilization of around 10 GB regardless of the number of files. Partial duplicate removal, or deduplication, results in a slight increase in storage usage, with 100,000 files consuming approximately 150 GB. This illustrates how deduplication methods can drastically reduce storage requirements compared to non-deduplication methods.

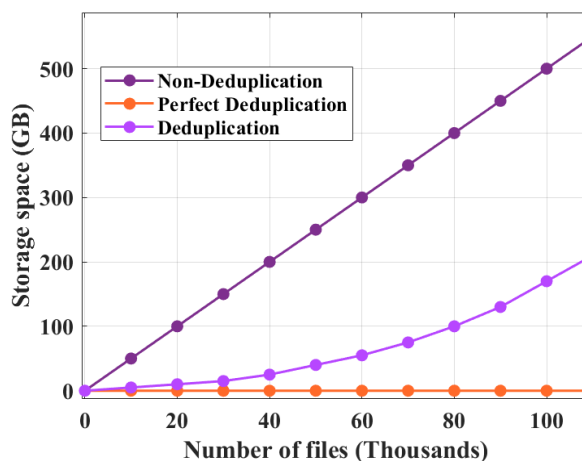


Fig. 7. Storage space requirements under non-deduplication, perfect deduplication, and partial deduplication scenarios.

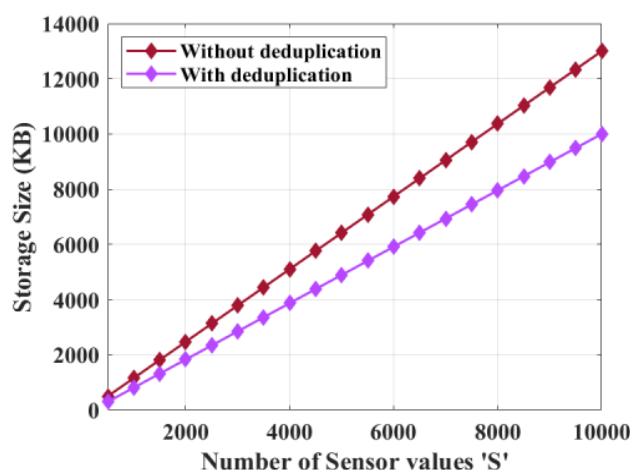


Fig. 8. Storage usage versus number of sensor values with and without deduplication.

Comparing situations with and without deduplication, Figure 8 illustrates the storage capacity requirements (in KB) as the number of sensor values ('S') increases. Storage increases linearly in the absence of deduplication, reaching about 13,000 KB for 10,000 sensor readings. The storage required increases more slowly with deduplication, reaching around 8,500 KB at the same time. Deduplication reduces storage usage from approximately 6,500 KB to around 4,250 KB at 5,000 sensor readings. By removing redundant data, deduplication effectively reduces storage needs and saves a substantial amount of storage space.

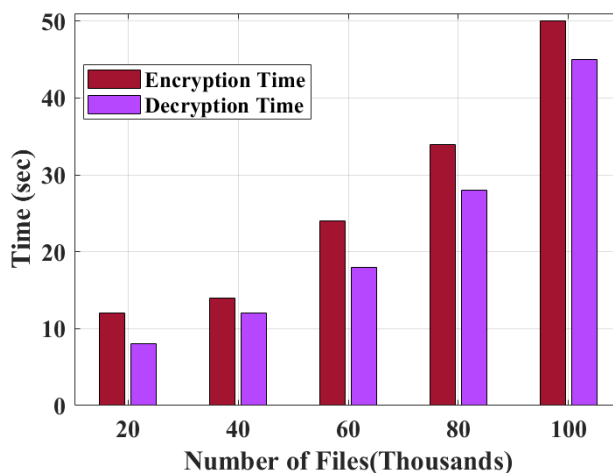


Fig. 9. Encryption and decryption time comparison across different file counts.

Comparing encryption and decryption timings for different file counts (in thousands) is shown in Figure 9. Red indicates a linear rise in encryption time, which goes from 10 seconds for 20,000 files to 50 seconds for 100,000 files. With a range of around 8 seconds for 20k files to 45 seconds for 100k files, the decryption time (purple) likewise increases linearly but stays continuously lower. While the growing duration shows that the computational overhead increases with the number of files, decryption is marginally more effective than encryption for all file sizes. This implies a balance in data management between computing performance and security.

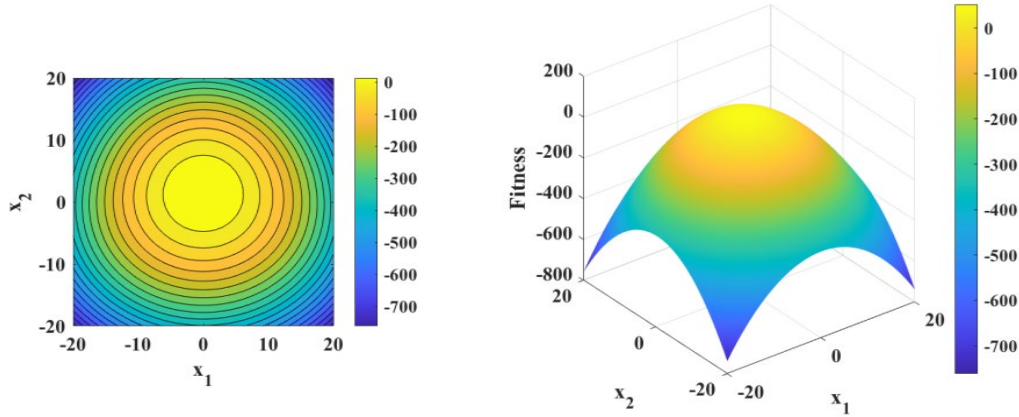


Fig. 10. Contour and 3D visualization of the fitness function landscape.

A fitness function's behavior across the variables x_1 x_2 is seen in Figure 10. Concentric levels with values ranging from 0 (yellow, peak) to -700 (blue, lowest) are displayed in the left contour plot. This fitness landscape may be seen in the right 3D figure, which indicates a parabolic peak at the origin ($x_1 = 0, x_2 = 0$) with a maximum value of 0. At the boundary, the function drops symmetrically in all directions, down to -700 ($x_1, x_2 = \pm 20$). With a single global maximum at its core, this image facilitates comprehension of the optimization issue.

4.3. Comparative analysis

The performance of the proposed strategy is compared to that of the established methods GH-EDA [16], CB-TTDD [17], EMCE [18], BGNBA-OCO [19], and EE-FDAS [20] using several metrics, such as energy consumption, throughput, average aggregation, packet loss ratio, storage cost, and encryption/decryption times. The analysis emphasizes the benefits of the suggested strategy for performance optimization.

Table 3. Comparison of energy consumption (EC) and average aggregation (AA) across different methods.

Methods	No of attributes					
	4		8		12	
	EC (μJ)	AA (%)	EC (μJ)	AA (%)	EC (μJ)	AA (%)
GH-EDA [16]	25	0.79	30	0.78	40	0.70
CB-TTDD [17]	28	0.82	35	0.75	45	0.68
EMCE [18]	20	0.85	25	0.80	35	0.72
BGNBA-OCO [19]	22	0.83	28	0.76	38	0.69
EE-FDAS [20]	18	0.88	22	0.80	30	0.74
Proposed	6	0.92	10	0.89	16	0.87

Table 3 displays the AA performance and EC of several techniques spanning attribute counts (4, 8, and 12). The EC values are 25, 30, and 40, while the AA values are 0.79, 0.78, and 0.70, per GH-EDA. EC is higher (28, 35, 45) while AA is somewhat lower (0.82, 0.75, 0.68), according to CB-TTDD. Better AA (0.85, 0.80, and 0.72) and decreased EC (20, 25, and 35) are the results of EMCE. Similarly, AA (0.83, 0.76, and 0.69) and EC (22, 28, and 38) are balanced by BGNBA-OCO. EE-FDAS demonstrates efficiency with AA (0.88, 0.80, 0.74), and EC (18, 22, 30). The proposed method performs effectively with low EC (6, 10, and 16) and high AA (0.92, 0.89, and 0.87).

Table 4. Deduplication ratio (%) of various methods under different node counts.

No of nodes Methods	De-duplication ratio (%)				
	60	120	180	240	300
GH-EDA [16]	85	86	85	84	84
CB-TTDD [17]	87	88	87	86	86
EMCE [18]	90	91	90	89	89
BGNBA-OCO [19]	93	94	93	92	92

EE-FDAS [20]	95	96	95	95	94
Proposed	99	99	98	98	98

De-duplication ratios (%) for six different approaches with different numbers of nodes (60, 120, 180, 240, and 300) are compared in Table 4. Although at lower concentrations, GH-EDA maintains stability with ratios of 85–86%. With a range of 87 to 88%, CB-TTDD performs somewhat better than the other options. With ratios ranging from 90% to 91%, EMCE outperforms BGNBA-OCO, which also improves, reaching ratios of 93% to 94%. EE-FDAS exhibits high efficiency, with values ranging from 94% to 96%. Superior de-duplication performance is ensured in all cases by the suggested strategy, which routinely achieves high ratios of 99% for 60 and 120 nodes and 98% for 180, 240, and 300 nodes.

Table 5. Throughput (T) and packet loss ratio (PLR) comparison for different attribute counts.

Methods	No of attributes					
	4		8		12	
	T (KB/s)	PLR (%)	T (KB/s)	PLR (%)	T (KB/s)	PLR (%)
GH-EDA [16]	120	15	140	14	160	13
CB-TTDD [17]	130	14	150	13	170	12
EMCE [18]	140	13	160	12	180	11
BGNBA-OCO [19]	145	12	165	11	185	10
EE-FDAS [20]	148	11	168	10	188	9
Proposed	150	10	170	9	230	8

In Table 5, the performance of various approaches is compared in terms of throughput (T) in KB/s and packet loss ratio (PLR) in % for variable amounts of attributes (4, 8, and 12). For all approaches, throughput improves and PLR falls as the number of attributes rises. Starting with a throughput of 120 KB/s at 4 attributes and a PLR of 15%, GH-EDA [16] improves to 160 KB/s at 12 attributes with a PLR of 13%. With PLRs of 10%, 9%, and 8%, the suggested approach consistently produces the highest throughput of 150 KB/s, 170 KB/s, and 230 KB/s for attributes 4, 8, and 12, respectively.

Table 6. Comparison of storage cost, encryption time, and decryption time among existing methods and the proposed approach.

Methods	Storage cost (bytes)	Encryption time (s)	Decryption time (s)
GH-EDA [16]	56	77	73
CB-TTDD [17]	154	79	75
EMCE [18]	146	67	64
BGNBA-OCO [19]	168	65	60
EE-FDAS [20]	27	59	55
Proposed	14	50	45

A comparison of several approaches is presented in Table 6, considering storage cost, encryption time, and decryption time. In comparison to other approaches such as GH-EDA (56 bytes) and CB-TTDD (154 bytes), the suggested method has the lowest storage cost at 14 bytes. The suggested approach is also the fastest in terms of encryption time, taking 50 seconds, which is less than GH-EDA's (77 seconds) and EE-FDAS's (59 seconds) requirements. GH-EDA (73 seconds) and CB-TTDD (75 seconds) are both slower than the suggested approach, which takes 45 seconds for decryption. The effectiveness of the suggested approach in reducing calculation and storage time is demonstrated by these findings.

4.4. Statistical analysis

The statistical study evaluates the performance of nodes at various fog layer levels. For healthcare data aggregation, statistical analysis uses metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to quantify the throughput of de-duplication to evaluate the performance of nodes at different Fog layer levels. The total amount of the error is shown by the average squared difference between the predicted and actual values. The square root of MSE, RMSE, provides a simpler statistic in the original data unit. At each layer, consistency and performance efficiency are assessed by considering the average, maximum, and lowest values across nodes in Fog Layers 1 and 2.

Table 7. Statistical analysis of ESC-SNGNet-HMO performance across two fog layers.

Node	Fog layer 1		Fog layer 2	
	MSE	RMSE	MSE	RMSE
Min	63.06	7.94	1.85	1.36
Max	63.32	7.96	5.75	2.40
Average	63.15	7.95	4.25	2.06

Reflecting the performance at various Fog layer levels, the statistical results are obtained from the model's output over two different Fog layers described in Table 7. For each layer, the model's optimal performance is shown by the least MSE and RMSE values. The minimal MSE for Fog Layer 1 is 63.06, while the RMSE is 7.94. The RMSE for Fog Layer 2 is 1.36, while the lowest MSE is 1.85. The maximum values display the model's largest reported inaccuracy at each layer. The maximal MSE for Fog Layer 1 is 63.32, while the RMSE is 7.96. The highest MSE for Fog Layer 2 is 5.75, while the RMSE is 2.40. These numbers illustrate the worst-case performance, where the model's predictions diverge further from the actual data. The overall model performance across several tests is represented by the average values of MSE and RMSE. The MSE for Fog Layer 1 is 63.15 with an RMSE of 7.95, and for Fog Layer 2, the MSE is 4.25 with an RMSE of only 2.06. These values indicate that the model performs better in Fog Layer 2's performance in Fog Layer 2 is displayed.

Statistical results obtained from ESC-SNGNNet-HMO also reveal that the system is capable of mitigating errors; the smaller the MSE and RMSE values, the better the model performs. The above statistics illustrate the recall of redundant data and energy usage from the operational deduplicated data in the Fog Server, regarding the availability of the aforementioned model. The enhanced accuracy of the metric and overall performance improvements in Fog Layer 2 also suggest that the use of SNGNN and HMO optimization has improved the accuracy of data collection and storage in this layer.

4.5. Ablation study

An ablation study demonstrates the importance of certain traits or components, as each is progressively removed and the resulting loss in performance is analyzed. The purpose of this study is to investigate the impact of four key components on the efficiency of the ESC-SNGNNet-HMO model for healthcare data aggregation in Internet of Things environments: ET (Energy Transformer), SNGNN (Similarity-Navigated Graph Neural Network), SDECC (Schoof's Dynamic Elliptic Curve Cryptography), and HMO (Human Memory Optimization). Comparing it with those models that are derived by leaving out at least one component of the whole model can provide a better understanding of the significance of each component towards different parameters of the system, such as memory usage, model complexity, response time, convergence time, and data transfer rate.

Table 8. Ablation study showing the effect of removing key components on system performance metrics.

Components	Memory usage (MB)	Model complexity (GFLOPs)	Response time (ms)	Convergence speed (nodes)	Data transfer rate (MB/s)
Without ET	47	463	459	500	37
Without SNGNN	38	34	43	50	51
Without SDECC	53	45	15	16	17
Without HMO	14	150	131	140	122
Full model (ET+SNGNN+SDECC+HMO)	50	51	53	45	51

Table 8 presents the performance of the ESC-SNGNNet-HMO model and its variants, as well as the impact of component elimination on significant variables. Removing SDECC increases memory usage to the highest level (53 MB), while removing HMO reduces it to the lowest (14 MB), indicating that SDECC contributes to memory efficiency rather than being the most memory-demanding component. Model complexity (GFLOPs) reaches its maximum (463 GFLOPs) when ET is removed, suggesting that the Energy Transformer helps reduce the model's computational load. Removing ET results in the highest response time (459 ms), indicating that the Energy Transformer has the most influence on response times. Removing HMO still results in a very high response time (131 ms). Without ET, convergence requires the most nodes (500), indicating that the Energy Transformer speeds up convergence and reduces the number of nodes needed. The entire model achieves balanced performance in all measures with values around the ideal range, although the data transfer rate (MB/s) is greatest when SNGNN is eliminated (51 MB/s). This is because removing the deduplication step results in higher data transmission.

4.6. Discussion

The proposed ESC-SNGNNet-HMO outperforms existing methods on certain parameters. It consistently beats alternatives like BGNBA-OCO (94%) and EE-FDAS (96%), with a deduplication ratio of up to 99% for lower node counts. The encryption and decryption times are reduced to 50 and 45 seconds, respectively, in comparison to EE-FDAS (59 and 55 seconds). Storage expenses are reduced to 14 bytes compared to EE-FDAS, which requires 27 bytes. With a throughput of 230 KB/s and a packet loss rate (PLR) of 8% for 12 characteristics, the method outperforms EE-FDAS (188 KB/s, 9% PLR). It balances lower computational overhead and deduplication efficiency with notable storage savings. In addition to highlighting optimization capabilities, the fitness function provides a solid option for safe data processing and deduplication in large-scale settings.

5. Conclusion

By emphasizing energy efficiency and data redundancy reduction, the ESC-SNGNNet-HMO scheme offers a novel approach to enhance data transmission in diverse healthcare settings. Through ET and SNGNN for deduplication, along with HMO for hyperparameter tuning, the system efficiently handles data aggregation, and deduplicated data is securely transmitted using SDECC. The framework achieves a deduplication ratio of up to 99%, significantly reduces energy consumption and storage costs (14 bytes), and maintains a low packet loss rate (8%). Although evaluation was performed using NS-2.35 simulations, the framework is compatible with real IoT healthcare deployments, as the architecture follows standard sensor, aggregator, fog, and cloud configurations. Key limitations of the proposed ESC-SNGNNet-HMO framework include increased computational demands with larger node counts, potential cryptographic overhead on resource-constrained devices, and the absence of experimental validation in real-world IoT healthcare deployments. Future work will address these challenges by enhancing scalability and computational efficiency, incorporating energy-efficient compression techniques for critical patient data without compromising quality, and conducting practical evaluations on real IoT healthcare datasets and prototype sensor networks. These steps aim to ensure that the framework remains both secure and resource-efficient while maintaining high deduplication performance and applicability across diverse healthcare environments.

All the Declarations and Statements

Author Contributions Statement

Aishwarya Shekhar – Conceptualization, Formal analysis, Software, Supervision, Project administration, Writing - Review & Editing.

Abdul Aleem – Methodology, Investigation, Resources, Validation, Data Curation, Writing - Original Draft.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

Authors do not have any conflict of interest to declare.

Funding Declaration

No funding is provided for the preparation of manuscript.

Data Availability Statement

Data sharing not applicable to this article.

Ethical Declaration

This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgments

Not applicable.

Declaration of Generative AI in Scholarly Writing

AI tools were used only for minor grammar checking and language refinement. All methodological development, analysis, and technical content were entirely carried out by the authors without AI assistance.

Abbreviation

The following abbreviations are used in this manuscript:

DR - Deduplication ratio
EC - Energy consumption
EE-FDAS - Energy-Efficient Fuzzy Data Aggregation System
EMCE - Enhanced Map Chaotic Encryption
ET - Energy Transformer
GH-EDA - Grid Hashing Efficient Data Aggregation
EHR - Electronic Health Records
HMO - Human Memory Optimization
IoT - Internet of Things
MSE - Mean Squared Error

OVFKS - Optimal Verified Fuzzy Keyword Search
RL-CO - Refraction Learning-based Chimp Optimization
RMSE - Root Mean Squared Error
SCDD - Secure Cloud Data Deduplication
SDECC - Schoof's Dynamic Elliptic Curve Cryptography
SNGNN - Similarity-Navigated Graph Neural Network
TL-CH - Tri-Level Chunk Hashing

References

- [1] M. Sharma, A. Tomar, and A. Hazra, "From Connectivity to Intelligence: The Game-Changing Role of AI and IoT in Industry 5.0," *IEEE Consum. Electron. Mag.* pp. 1–7, Sep. 27 2024. DOI: 10.1109/MCE.2024.3470340
- [2] S. Rani and G. Srivastava, "Secure hierarchical fog computing-based architecture for industry 5.0 using an attribute-based encryption scheme," *Expert Syst. Appl.*, vol. 235, p. 121180, Jan. 1 2024. DOI: 10.1016/j.eswa.2023.121180
- [3] P. K. Donta, I. Murturi, V. Casamayor Pujol, B. Sedlak, and S. Dustdar, "Exploring the potential of distributed computing continuum systems," *Computers (Basel)*, vol. 12, no. 10, p. 198, Oct. 2 2023. DOI: 10.3390/computers12100198
- [4] P. Palanisamy, A. Padmanabhan, A. Ramasamy, and S. Subramaniam, "Remote patient activity monitoring system by integrating IoT sensors and artificial intelligence techniques," *Sensors (Basel)*, vol. 23, no. 13, p. 5869, Jun. 25 2023. DOI: 10.3390/s23135869
- [5] O. L. Băjenaru, L. Băjenaru, M. Ianculescu, V. Ş. Constantin, A. M. Guşatu, and C. R. Nuţă, "Geriatric Healthcare Supported by Decision-Making Tools Integrated into Digital Health Solutions," *Electronics (Basel)*, vol. 13, no. 17, p. 3440, Aug. 30 2024. DOI: 10.3390/electronics13173440
- [6] A. S. Abdul-Qawy, N. A. Alduais, A. M. Saad, M. A. Taher, A. B. Nasser, S. A. Saleh, et al., "An enhanced energy efficient protocol for large-scale IoT-based heterogeneous WSNs," *Sci. Afr.*, vol. 21, p. e01807, 2023. DOI: 10.1016/j.sciaf.2023.e01807
- [7] W. Deng, X. Li, J. Xu, W. Li, G. Zhu, and H. Zhao, "BFKD: Blockchain-based federated knowledge distillation for aviation Internet of Things," *IEEE Trans. Reliab.* pp. 1–14, 2024. DOI: 10.1109/TR.2024.3474710
- [8] B. Bhasker and S. Murali, "An Energy-Efficient Cluster-based data aggregation for agriculture irrigation management system using wireless sensor networks," *Sustain. Energy Technol. Assess.*, vol. 65, p. 103771, 2024. DOI: 10.1016/j.seta.2024.103771
- [9] M. Kumar, M. Sethi, S. Rani, D. K. Sah, S. A. AlQahtani, and M. S. Al-Rakhami, "Secure Data Aggregation Based on End-to-End Homomorphic Encryption in IoT-Based Wireless Sensor Networks," *Sensors (Basel)*, vol. 23, no. 13, p. 6181, Jul. 6 2023. DOI: 10.3390/s23136181
- [10] G. Tan, H. Yuan, H. Hu, S. Zhou, and Z. Zhang, "A Framework of Decentralized Federated Learning With Soft Clustering and 1-Bit Compressed Sensing for Vehicular Networks," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 23617–23629, 2024. DOI: 10.1109/IJOT.2024.3385913
- [11] S. U. Muthunagai and R. Anitha, "CTS-IIoT: Computation of time series data during index based de-duplication of industrial IoT (IIoT) data in cloud environment," *Wirel. Pers. Commun.*, vol. 129, no. 1, pp. 433–453, 2023. DOI: 10.1007/s11277-022-10105-5
- [12] M. A. Aleisa, N. Beloff, and M. White, "Implementing AIRM: A new AI recruiting model for the Saudi Arabia labour market," *J. Innov. Entrep.*, vol. 12, no. 1, p. 59, 2023. DOI: 10.1186/s13731-023-00324-w
- [13] E. Zeydan, S. S. Arslan, and M. Liyanage, "Managing Distributed Machine Learning Lifecycle for Healthcare Data in the Cloud," *IEEE Access*, vol. 12, pp. 115750–115774, 2024. DOI: 10.1109/ACCESS.2024.3443520
- [14] J. Ofulue and M. Benyoucef, "Data monetization: Insights from a technology-enabled literature review and research agenda," *Management Review Quarterly*, vol. 74, no. 2, pp. 521–565, 2024. DOI: 10.1007/s11301-022-00309-1
- [15] Tang, X., Guo, C., Choo, K.-K. R., Jiang, X., & Liu, Y. (2024). A secure and lightweight cloud data deduplication scheme with efficient access control and key management. *Computer Communications*, 222, 209–219. DOI: 10.1016/j.comcom.2024.05.003
- [16] S. M. Altowaijri, "Efficient Data Aggregation and Duplicate Removal using Grid-Based Hashing in cloud-assisted Industrial IoT," *IEEE Access*, vol. 12, pp. 145350–145365, 2024. DOI: 10.1109/ACCESS.2024.3471952
- [17] M. Akbar, I. Ahmad, M. Mirza, M. Ali, and P. Barmavatu, "Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach," *Cluster Comput.*, vol. 27, no. 3, pp. 3683–3702, 2024. DOI: 10.1007/s10586-023-04171-y
- [18] T. Benil and J. Jasper, "Blockchain based secure medical data outsourcing with data deduplication in cloud environment," *Comput. Commun.*, vol. 209, pp. 1–3, 2023. DOI: 10.1016/j.comcom.2023.06.013
- [19] M. Pavithra, M. Prakash, and V. Vennila, "BGNBA-OCO based privacy preserving attribute based access control with data duplication for secure storage in cloud," *J. Cloud Comput. (Heidelb.)*, vol. 13, no. 1, p. 8, 2024. DOI: 10.1186/s13677-023-00544-1
- [20] M. N. Khan, W. Cao, Z. Tang, A. Ullah, and W. Pan, "Energy-Efficient De-Duplication Mechanism for Healthcare Data Aggregation in IoT," *Future Internet*, vol. 16, no. 2, p. 66, 2024. DOI: 10.3390/fi16020066
- [21] Z. Wang, W. Gao, M. Yang, and R. Hao, "Enabling Secure Data sharing with data deduplication and sensitive information hiding in cloud-assisted Electronic Medical Systems," *Cluster Comput.*, vol. 26, no. 6, pp. 1–16, Oct. 28 2022.
- [22] A. Javadpour, A. M. Abadi, S. Rezaei, M. Zomorodian, and A. S. Rostami, "Improving load balancing for data-duplication in big data cloud computing networks," *Cluster Comput.*, vol. 25, no. 4, pp. 2613–2631, 2022. DOI: 10.1007/s10586-021-03312-5
- [23] M. Thottipalayam Andavan, M. Parameswari, N. Subramanian, and N. Vairaperumal, "A novel model for enhancing cloud security and data deduplication using fuzzy and refraction learning based chimp optimization," *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 3, pp. 1025–1038, 2024. DOI: 10.1007/s13042-023-01953-z

- [24] E. G. Nascimento, T. A. de Melo, and D. M. Moreira, "A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy," *Energy*, vol. 278, p. 127678, 2023. DOI: 10.1016/j.energy.2023.127678
- [25] M. Zou, Z. Gan, R. Cao, C. Guan, and S. Leng, "Similarity-navigated graph neural networks for node classification," *Inf. Sci.*, vol. 633, pp. 41–69, 2023. DOI: 10.1016/j.ins.2023.03.057
- [26] D. Zhu, S. Wang, C. Zhou, S. Yan, and J. Xue, "Human memory optimization algorithm: A memory-inspired optimizer for global optimization problems," *Expert Syst. Appl.*, vol. 237, p. 121597, 2024. DOI: 10.1016/j.eswa.2023.121597
- [27] D. Maimuț and A. C. Matei, "Speeding-Up Elliptic Curve Cryptography Algorithms," *Mathematics*, vol. 10, no. 19, p. 3676, 2022. DOI: 10.3390/math10193676

Authors' Profiles



Aishwarya Shekhar is working as an Assistant Professor in the Department of Computer Science and Engineering at Sandip University, Sijoul, Madhubani. He is currently pursuing a Ph.D. in Computer Science and Engineering from Galgotias University, with an expected completion in 2025. His academic journey has been marked by a consistent pursuit of excellence, as evidenced by securing an A+ grade in his Ph.D. coursework exams. Mr. Shekhar also completed a Post Graduate Diploma in Management (PGDM) in IT from SVKM's Narsee Monjee Institute of Management Studies in 2022. He holds a Master of Technology degree in Computer Science and Engineering from Manipal University Jaipur, obtained in 2015, and a Bachelor of Technology in Computer Science and Engineering from UPTU, where he graduated with first division honors in 2013. Professionally, Mr. Shekhar began his career at AWC Software Pvt. Ltd, where he served as a Software Developer. He then transitioned to Velocis Systems Pvt. Ltd as a Senior Software Developer. Currently, he is the Academic Coordinator for the B.Tech and M.Tech courses at Sandip University, Madhubani. He has published more than 20 research papers in national and international journals, including five Scopus-indexed papers and three IEEE Conference Papers, which are also indexed in Scopus. Additionally, he has co-authored a book. Mr. Shekhar has six years of experience in the IT industry and three years in academia. He is a member of IEEE, ACM, and IAENG. His research interests include IoT, Cloud Computing, AI, and Machine Learning.



Abdul Aleem is a Professor in the School of Computing Science and Engineering (SCSE), at Galgotias University. He has done his B. Tech. (CSE) from U.P. Technical University, Lucknow. He has done his M. Tech. (Software Engg.) from MNNIT Allahabad. He was bestowed Gold Medal for being the topper of his batch in the master's program. He has obtained Ph.D. (CSE) from MNNIT Allahabad in the area of Data Mining, Machine Learning, and Educational Systems. He has been thrice awarded as Best/Exemplary Faculty, five times conferred the Best Paper award, and two times awarded for coding Excellence. He possesses 13+ years of blended experience including 6+ years in top software industries and 7+ years in premium educational institutions as a value-added academician, software developer, lead engineer, researcher, administrator, and intellectual professional. His research interests include digital image processing, optimization techniques, machine learning and artificial intelligence. He has published more than 30 research papers in reputed conferences and journals of distinguished indexing.

How to cite this paper:

Aishwarya Shekhar, Abdul Aleem, "Similarity-Navigated Graph Neural Network: Energy-Efficient De-Duplication for Healthcare Data Aggregation in IoT Environments", *International Journal of Computer Network and Information Security(IJCNIS)*, Vol.18, No.4, pp. 46-63, 2026. DOI:10.5815/ijcnis.2026.04.03